新闻 App 用户行为分析的实现

摘　要：本文对新闻 App 用户行为数据采集方式、数据分析方法进行了研究，提出了一种适用于新闻 App 用户行为高效分析的方法。该方法使用 splunk 进行后端数据分析，通过对列方式解决用户行为数据高并发问题，保证了 App 的高可用性。文中叙述的方法已经在实际使用的 App 中进行了部署，并经受了长时间的运行考验，可为类似新闻阅读类应用提供现实参考。
关键词：用户行为分析；数据采集；数据分析；队列；splunk
中图分类号：TP311.5；TP316　　文献标识码：A
文章编号：1671-0134（2019）03-039-05　　DOI：10.19483/j.cnki.11-4653/n.2019.03.006

文 / 赵强　彭玮

1. 用户行为分析意义

用户行为，简单来说就是用户在网站或 App 上进行操作而产生的一系列行为。

在获取网站或者 App 等平台用户访问基本数据后，对相关数据进行统计分析，从中发现用户的行为习惯和潜在需求，有针对地解决业务相关问题、提高用户体验，并为决策提供依据的过程我们称之为用户行为分析。

那么，为什么要进行用户行为分析？截至 2018 年 12 月底，我国网站数量为 523 万个，App 在架数量为 449 万款，每个领域都拥有成百上千的网站和移动 App 产品，竞争异常激烈。在这种环境下，如果企业能做好精细化的用户行为分析、找准问题所在，并有效提出改进方向，既能避免资源浪费，又有助于优化产品质量，提升公司竞争力。

用户行为分析通常应用于以下场景：

（1）运营分析，用户来源渠道统计，有助于判断哪个渠道有利于拉拢新用户。

（2）产品改版分析，通过 AB 测试或其他方法来分析不同 App 版本下的用户行为，以此来判断哪个版本会受用户喜爱，最终上线哪个版本。

（3）预测分析，合理构建算法模型，通过训练模型、验证模型，最终使用模型对用户关键行为进行预测，提前洞察可能出现流失的环节，提早干预，从而降低流失率。

（4）采用基于用户的协同过滤算法进行产品或内容推荐，即为 A 用户推荐相似用户 B 用户感兴趣的产品或者内容。用户画像构建是前提，用户行为越多，推荐越精准，用户的使用率和使用时长也会优化。

2. 新闻 App 用户行为分析

据 2018 年 2 月底发布的《第 43 次中国互联网络发展状况统计报告》显示，网络新闻占据中国网民各类手机应用的前三位，截至 2018 年 12 月底，手机网络新闻用户达到 6.53 亿，占手机网民的 79.9%。由此可见，通过移动设备获取新闻资讯已经成为非常重要的渠道。

新闻媒体，从纸媒到门户再到移动端。传统媒体多是以文字和图片呈现，而新媒体不仅可以通过文字、图片，还可以通过图集、视频等多元化方式展现新闻内容，对用户的冲击力更大、更直观。在新闻内容传播的同时，用户还可以进行评论、点赞、转发，增加了互动社交性也扩大了传播性。新媒体的时效性、传播迅速也是传统媒体远远比不上的。当用户不仅仅是看新闻而是使用一个新闻资讯产品时，我们要在考虑满足用户对新闻内容需求的同时也要考虑到用户的体验。为了更好地了解我们新闻 App 产品的使用情况和不足，本文将就新闻类 App 的用户行为分析技术做些分析和总结。

	2018.12		2017.12		
应用	用户规模（万）	手机网民使用率	用户规模（万）	手机网民使用率	年增长率
手机即时通信	78029	95.5%	69359	92.2%	12.5%
手机搜索	65396	80.0%	62398	82.9%	4.8%
手机网络新闻	65286	79.9%	61959	82.3%	5.4%
手机网络购物	59191	72.5%	50563	67.2%	17.1%
手机网络视频	58958	72.2%	54857	72.9%	7.5%
手机网上支付	58339	71.4%	52703	70.0%	10.7%

图 1　手机网民各类手机互联网应用使用率图

3. 用户行为数据采集

用户行为数据的采集方式直接决定了数据源的质量，是用户行为数据分析的基础。数据的采集通常采用“埋点”方式。埋点又可分为客户端埋点（前端埋点）和服务器埋点（后端埋点）。

3.1 客户端埋点

客户端埋点还可分为代码埋点和可视化埋点。

3.1.1 代码埋点

前端的代码埋点，顾名思义是在产品开发阶段，依据 PM 要求的数据需求文档，前端开发人员在每个需要采集的数据点写入代码。在用户每次前端操作时能够触发埋点上传数据。

3.1.2 可视化埋点

开发加入“无埋点”的采集代码，能够对网页或者

App 上所有的可交互事件元素进行解析并监测，当有用户操作行为（交互事件）发生时，即对此事件进行采集、上报。无埋点并不是不用写入任何代码，而是通过代码将所有事件元素解析后，以可视化的方式呈现，让 PM、运营经理等可以根据需要自行手动选取、标定。为了与开发逐一进行代码写入的方式进行区分，被称作可视化埋点。可视化埋点通常通过第三方工具实现。

3.2 服务器端埋点

服务器端埋点也成为后端埋点，指的是开发在服务器端写入代码，采集了前后端的交互事件数据以及存储与业务服务器中的业务数据。

三种方式对比如下表 1 所示：

表 1　三种埋点方式对比表

	可视化埋点	前端代码埋点	后端代码埋点
准确度	低	中	高
前期工作量	低	高	高
后期工作量	高	中	低
优点	部署简单；便于运营和产品配置	准确度高；自定义强	准确度高；自定义强
缺点	不灵活；网络传输压力大	工作量大	工作量大
第三方产品	Heap Analytics、Growing IO	Google Analytics、TalkingData	

4. 分析数据

4.1 分析指标

行为数据分析，通常要靠指标分析结果，选取合适的指标，才能有效判断用户的使用情况和 App 的问题所在。下面介绍几种常用的指标。

4.1.1 新增用户

新增用户指的是 App 被下载安装后第一次启动 App 的用户，用于衡量推广的效果。通常按照时间和渠道来源分类。

4.1.2 活跃用户

活跃用户指的是在一定统计周期内打开 App 的用户数，一般用来衡量 App 的运营现状。

4.1.3 用户留存率

用户留存率是指在某一个统计时段的新增用户数中经过了一段时间后仍打开这个 App 的用户比例，包括次日留存、7 日留存（如今天新增用户数在第 7 日再次打开 App 的比例，14 日和 30 日留存以此类推）、14 日留存、30 日留存。这个指标是验证你的 App 对用户是否具有吸引力。

4.1.4 启动次数

统计某一时段用户打开 App 的次数。

4.1.5 使用时长

使用时长是指在统计周期内所有用户从打开 App 到关闭 App 的总时长。这个指标考核的是你的 App 用户粘性高不高，也反映了 App 的产品质量高低，使用时长一般会结合启动次数一起分析。

4.1.6 用户画像分析

有了用户数据，再做用户画像分析会更加容易。用户画像是对人口属性的特征分析、用兴趣分析、用户行为分析等。用户画像可以帮助 App 逐渐实现精准化营销，直接进行 App 与指定用户之间的点对点交互。

4.2 分析工具使用

用户行为数据分析工具通常有以下几种：

4.2.1 脚本统计

Python、R 语言是常用的数据分析语言，优点是成本不会太高，缺点是开发周期长、学习成本高，而且运行效率可能不高。

4.2.2 第三方统计

Talking Data、Growing IO、Google Analytics 等工具，都可以用于数据分析，优点是方便快捷，缺点是成本会比较高。

4.2.3 日志分析工具

日志分析工具，如 splunk、ELK，可以分析各种数据，而且有统计功能，优点是可以更灵活地运用于自身的业务。

5. 实现方法

本文实现了新闻 App 的用户行为分析，通过前端埋点方式采集客户端用户行为数据，将数据传入后端，后端将行为数据放入队列，splunk 通过应用接收队列数据后定时进行分析，并将统计结果放回队列中，程序取出队列数据放入 MongoDB 数据库，最终页面展示在业务系统中，供业务人员使用。流程如下图 2 所示：

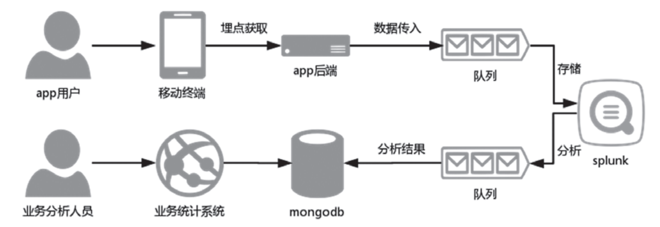


图 2　App 用户行为分析流程

5.1 前端埋点

获取用户行为数据采用前端埋点的方式，根据分析需求，规划埋点位置，定义埋点字段，可获取真实的用户行为数据，埋点上传字段如下表 2 所示：

表 2　前端埋点字段

字段	含义
operateId	埋点采集动作类型
time	操作时间
ip	设备 ip 地址
position	设备位置
clientinfo	手机型号
contentType	操作信息类型
contentId	文章 id
title	文章标题
deviceId	手机 imer 号
appSouce	App 名称
msgword	搜搜
id	uuid
createtime	入库时间（后台入库时插入）
userId	已登录用户上传用户 id
channel	App 渠道名
version	App 版本
sysVersion	设备系统版本
deviceScreen	设备分辨率
netType	网络类型
platform	手机平台（Android/ios）
deviceToken	ios 传
coordinates	上传设备经纬度

表 3 为埋点采集动作类型（字段：operateId）。

表 3　埋点采集动作类型

动作	数值	对象
注册	1	用户
登录	2	用户
打开	3	新闻、栏目（要闻等）、app
关闭	4	新闻、栏目（要闻等）、app
收藏	5	新闻
赞	6	新闻
分享	7	新闻
订阅	8	站点
评论	9	新闻
搜索	10	关键词
安装	11	App

表 4 为操作信息类型（字段：contentType）。

表 4　操作信息类型

动作	数值
栏目	1
专栏	2
普通文章	3
图片	4
视频	5
直播间	6
专题	7
华媒	8
评论	9
用户相关	10
App	11

5. 2 数据可靠传输机制

由于用户行为数据并发量大且频率高，因此，我们需要采取合适的方式保证数据的可靠传输并且减少对后端的压力。这里我们采用消息队列方式进行数据传输，消息队列指的是在消息传输过程中保存消息的容器，埋点获取的用户行为信息，通过 App 后端传入队列 rabbitmq 中，采用消息队列临时存储数据，有以下好处：

5.2.1 异步处理，减少请求响应时间

一次用户行为， App 前端获取埋点数据后会向后端进行请求，后端接收数据后将其存入预先规定好的存储设备中。然而，这样 App 后端就会承受较大压力，一是前端频繁的请求，二是存储数据请求响应的过程。使用队列方式，可以减少请求响应时间，将数据直接放入队列中，然后让消费者再去存储数据，异步处理，减轻了后端压力。

5.2.2 应用解耦，不影响 App 的正常运行

当 App 后端要去存储用户行为数据时，如果存储设备出现无法访问等问题，那么 App 的正常运行可能就会受到影响。这时如果中间加入队列方式，App 后端将数据传入队列，就不会影响到 App，数据只会堆积到队列中，但也能确保用户行为数据的完整性。

综上，我们采用了队列方式存放及管理埋点数据，以降低前端数据采集与后端数据处理系统之间的耦合性，提升系统运行效率。

5.3 splunk 应用接收 mq 数据

队列中的用户行为数据需要由消费者取走，通常可以用程序方式，消费队列数据将其放入数据库中，再用脚本对数据进行分析。然而，我们会面临如下问题：首先原始的用户行为数据数据量非常庞大，普通数据库不适宜存储大数据，二是如果用脚本操作数据库，数据库会承受较大压力，分析数据的速度也会受到影响。这里我们选择采用 splunk 来接收存储数据。

Splunk 是一个典型的大数据处理工具，面向机器数据的全文搜索引擎，是一个一体化的平台：数据采集 – 存储 – 分析 – 可视化。Splunk 的应用 AMQP Messaging 经过配置可直接接收队列数据，存于 splunk 中，splunk 的专用搜索语言 SPL（searchprocessinglanguage）语法简单，类似 sql，可以直接进行分析，效率高速度快。与 splunk 类似的工具还有 ELK，ELK 是 ElasticSearch，Logstash，Kibana 的缩写，分别提供搜索，数据接入和可视化功能，ElasticSearch 是一个基于 Lucene 的开源搜索服务。使用 splunk 有如下优势：

（1）数据导入简单，splunk 提供各种应用可以直接接收不同数据，适配性强，而 ELK 需要 filebeats 或者 logstash 接收，接收后再传入 ES，配置过程较为繁琐。如图 3 所示，splunk 可以通过应用接收 MySQL 数据、Citrixnetscaler 数据等等，经过配置即可导入。其中 amqp_ta 即为接收队列数据的应用，部分配置如图 4 所示。图 5 为导入的用户行为数据。

Name ⇕	文件夹名称 ⇕
SplunkForwarder	SplunkForwarder
SplunkLightForwarder	SplunkLightForwarder
Splunk Add-on for Citrix NetScaler	Splunk_TA_citrix-netscaler
Splunk Add-on for MySQL	Splunk_TA_mysql
Palo Alto Networks Add-on	Splunk_TA_paloalto
Palo Alto Networks	SplunkforPaloAltoNetworks
Log Event Alert Action	alert_logevent
Webhook Alert Action	alert_webhook
amqp_ta	amqp_ta

图 3　splunk 安装应用图

Virtual Host
/
Broker Virtual Host
Use SSL ?
Message Consumer Settings
Queue Name
appUserActionSplunk
Name of queue to consume from
Exchange Name
app_useraction
Exchange name.Leave empty to use the Default Exchange.
Routing Pattern
fanout

图 4　splunk 配置队列数据导入

```
19/03/17      { [-]
18:41:56.000    appSouce: chinanews
                channel: huawei
                clientInfo: VKY-AL00
                contentId: 8782157
                contentType: 5
                coordinates: 32.125086,118.93862
                createtime: 2019-03-17 18:41:56
                deviceId: 865582034956038
                deviceScreen: 1080*1920
                deviceToken:
                id: 10072457c0c72c244be284b0447ebd279f7a
                ip: 117.62.250.126
                msgword:
                netType: wifi
                operateId: 4
                platform: android
                position: 中国|江苏省|南京市|栖霞区,鼎山东路
                pubtime: 2019-03-17 12:20:20
                sysVersion: 8.0.0
                time: 1552819289178
                title: 成都七中实验学校食堂食材粉条不合格 网传照片或系人为造假
                userId: 5c204d1c263ec56ba09abdc0
                version: 6.5.6
                videoId:
                videoType: tg
              }
              显示为原始文本
              channel = huawei   host = mq2   sourcetype = json:custom
```


图 5　用户行为数据事件

导入 elasticsearch 的数据要通过 logstash 处理为符合要求的数据格式，通过配置 logstash/config/logstash.conf 文件中 input、filter、output 模块，如下图 6 配置 input 模块和图 7 配置 output 模块，最终导入 elasticsearch，图 8 即为 kibana 查看导入 es 的 json 格式数据。

```
input {
        beats  {
                port => 5044
        }
        rabbitmq {
                host => "[redacted]57"
                port=> 5672
                user => "[redacted]"
                password => "[redacted]"
                queue => "apptoelk"
                durable => true
                codec => "json"
                type => "mq"
        }
}
```

图 6　logstash 配置 input 模块

```
output {
if [type] == "mq" {
        elasticsearch {
                hosts => ["127.0.0.1:9200"]
                index => "appuser_access_%{+YYYY.MM.dd}"
                template => "/usr/local/logstash-5.5.1/config
/es_template/mq.json"
                template_name => "mq*"
                template_overwrite => true
                        }
}
```

图 7　logstash 配置 output 模块

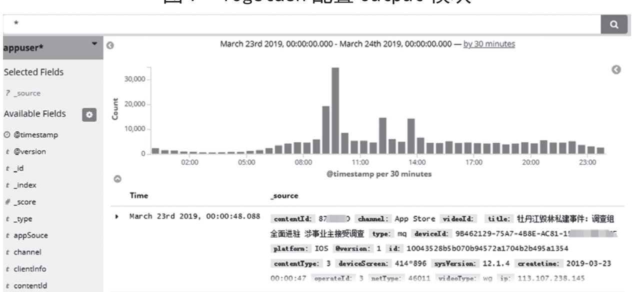


图 8　导入 elasticsearch 后 kibana 可视化图

（2）字段识别容易，splunk 支持自动抽取字段，搜索时也会动态抽取新的字段，而这是 elastic 不支持的。

（3）数据分析和处理上，ElasticSearch 使用 Search API 来实现，而 splunk 提供较强大的 SPL，语法简洁，非常易用，以查看某天的新闻阅读量排行为例，splunk 的 SPL 语句如下：

host=mq2

| search appSouce=chinanews operateId=3 contentType=3

| stats count as pv by contentId | sort –pv

Python 调用 ES 的脚本如下图 9 所示，比 SPL 略复杂，如果需要更复杂的计算，更会凸显 SPL 的优势。



```
es = Elasticsearch([{'host':[redacted]','port':9200}])
query = {
        "query": {
                "bool": {
                        "must": [{
                                "term": {
                                        "appSouce": "chinanews",
                                }
                        },
                        {
                                "term": {
                                        "contentType": "3"
                                }
                        },
                        {
                                "term": {
                                        "operateId": "3"
                                        }
                        }]
                }
        },
        "aggs":{
                "pv": {
                        "terms":{
                                "field":"contentId",
                                "size" : 100
                        }
                }
        }
}
searchresult = es.search(index='appuser_access_2019.03.23',  body=query)
for newspv in searchresult['aggregations']['pv']['buckets']:
    print newspv
```

图 9　Python 调用 ES

contentId	pv
8788405	607
8788273	525
8788431	443
8788244	435
8788524	420
8788173	386
8788409	377
8788028	326
8788482	304
8788403	263

图 10　splunk 分析结果图

通过图 10splunk 结果和图 11python 调用 ES 结果前十名对比，可以看出，分析结果稍有偏差，从 kibana 查看访问量最高的 contentId8788405 如图 12 所示，可以看出结果应该和 splunk 统计的相同均为 607 才对，但是聚

合运算过后，结果就不准确了。查看 ES 官方文档可以了解到 api 提供的聚合运算确实存在一定误差。因此语法简单易懂，计算准确的 splunk 就更有优势了。

```
{u'key': u'8788405', u'doc_count': 732}
{u'key': u'8788273', u'doc_count': 540}
{u'key': u'8788524', u'doc_count': 536}
{u'key': u'8788431', u'doc_count': 523}
{u'key': u'8788409', u'doc_count': 430}
{u'key': u'8788482', u'doc_count': 354}
{u'key': u'8788403', u'doc_count': 317}
{u'key': u'8788658', u'doc_count': 281}
{u'key': u'8788173', u'doc_count': 264}
{u'key': u'8788455', u'doc_count': 257}
```

图 11　Python 调用 ES 分析结果图

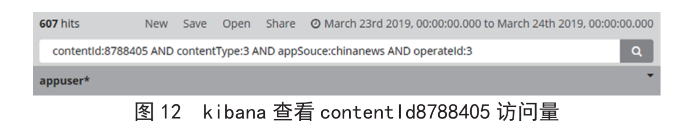


图 12　kibana 查看 contentId8788405 访问量

（4）可视化方面，ELK 用的是 kibana，splunk 直接在平台上集成了非常方便的数据可视化和仪表盘功能，通过简单配置就可以进行可视化分析。

5.4 splunk 定时统计输出

Splunk 的告警功能能够满足对数据进行定时统计分析的需求。如下图 13 所示，每天 0 点执行 SPL 语句，语句实现了统计昨日访问量前 100 名的正文稿件，结果数量大于 0 时，触发执行 Python 脚本 mq.py，Python 脚本实现将统计结果处理后放到队列。



图 13　splunk 定时统计分析

5.5 页面展示

Java 程序取上一步队列里的结果，放入 mongodb 数据库中。业务统计系统将结果展示在页面中，某天的正文访问量排行部分内容统计如下图 14 所示。

正文　直播　　1小时　3小时　全天

序号	发布时间	新闻标题	访问量
1	2019-03-15 16:50:07	李克强开诚布公答中外记者问 谈经济热题解民生难点	826
2	2019-03-16 17:58:39	努尔·白克力严重违纪违法被开除党籍和公职	769
3	2019-03-15 22:46:29	3·15晚会曝光了这些消费陷阱，别再被坑了	697
4	2019-03-16 00:08:42	香椿自由取代车厘子自由 最贵的1斤价格超200元	443
5	2019-03-16 01:45:25	3·15晚会落幕，这些企业被就地查处！	227
6	2019-03-16 11:51:00	360度解析卡纳瓦罗！他可能是国足史上最帅的主帅了……	183
7	2019-03-15 21:37:13	医生说过哪些令你震惊的话？了解一下	180

图 14　业务统计系统部分统计结果

结语

综上，splunk 为用户提供了一个存储和处理数据的平台，以最简单的方式将数据接入平台，最快的速度计算数据，让业务人员根据自身需求，利用平台上的数据解决自己实际业务中的问题。这样的方式可以在更多的行业和领域进行复制。将 splunk 应用于新闻 App 用户行为分析的实现，帮助我们很好地贴合实际业务进行分析，更深层次认识产品现状，对 App 的运营开发、改善优化具有指导意义。媒

参考文献

[1]http://www.woshipm.com/user-research/588417.html 应用场景 .

[2]http://www.woshipm.com/data-analysis/574447.html.

[3]《第 43 次中国互联网络发展状况统计报告》http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201902/P020190228510533388308.pdf.

[4]http://www.kejilie.com/chanpin100/article/MzaYJz.html.

[5]https://www.ichdata.com/app-user-behavior-monitoring.html.

[6]https://blog.csdn.net/kingcat666/article/details/78660535.

[7] http://www.huodonghezi.com/news-764.html.

[8]https://splunkbase.splunk.com/app/1812/.

[9]https://blog.51cto.com/splunkchina/1948105<Splunk 和 ElasticSearch 深度对比解析 >.

[10]https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations-metrics-cardinality-aggregation.html.

[11]https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations-bucket-terms-aggregation.html#_shard_size_3.

[12]http://beta.dooland.com/index.php?s=/magazine/article/id/954935.html.

（作者单位：中国新闻社）